A Comparative Analysis of NLP Algorithms and Their Recent Advances

Chelsea Chauhan, Komal Saxena, Nidhi Sindhwani

[1]Student, [2]Associate Professor, [3]Assistant Professor
[1]Amity Institute of Information Technology,
[1]Amity University, Noida, Uttar Pradesh, India
[1]chelsea.chauhan@s.amity.edu, [2]ksaxena1@amity.edu, [3]nsindhwani@amity.edu

Abstract - Natural Language is spoken and written by humans and thus, it is generated in an enormous amount in this world. Though it may have a big deal of knowledge inside it, but because of its large volume, it is becoming very hard day by day to propagate the gained knowledge by a person/human in a limited time span. The natural language processing is the answer for this job which provides the fruitful results using big amount of data, with good accuracy just like a human being. The following research represents the techniques of natural language processing, Natural Language Processing classification and fields where it is being used. Additionally, it covers the recent trends and advances made so far in the field and the recent algorithms being developed and used by well-known organizations. The research presents the comparative study of various machine learning and natural language processing algorithms on Language Detection which is an application of natural language processing and finally, the conclusion is presented based on the study.

Index Terms - Machine Learning(ML), Natural Language Processing(NLP), Artificial Intelligence(AI), Natural Language Understanding(NLU), Multinomial Naïve Bayes(MNB)

I. INTRODUCTION

Natural Language Processing is the potentiality of a machine or computer program to process and comprehend the human language or otherwise referred to as the natural language, as it is written or spoken by the humans. The parent branch of computer science is artificial intelligence that is the human like intelligence of machines which comes from the ability of a machine to learn from experiences called machine learning. NLP and several other types of machine-based intelligence fall under the umbrella of AI (Figure 1).

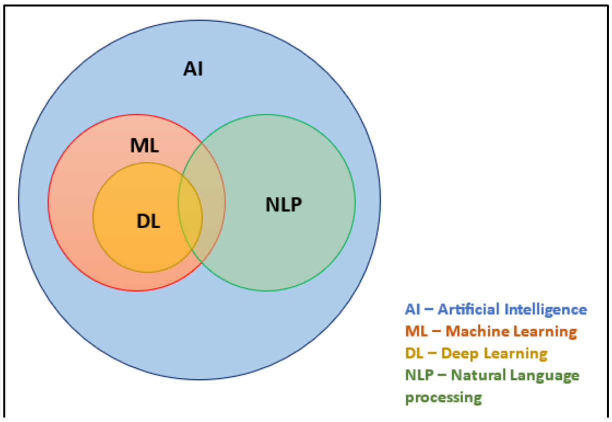

AI
ML
DL
NLP
AI – Artificial Intelligence
ML – Machine Learning
DL – Deep Learning
NLP – Natural Language processing

Fig. 1 Branches of AI

Natural language processing came into being around 1940s, after the World War II, when people started to acknowledge the need for language translations and anticipated for a machine to be able to do it automatically. The Turing test was proposed by Alan Turing in his renowned 1950 article "*Computing Machinery and Intelligence*," which established it as a standard of intelligence. In 1957, Noam Chomsky's *Syntactic Structures*, a rule-based system of syntactic structures, revolutionised linguistics [1]. One of the several other reasons of NLP coming into existence was to content the human hope of communication with machines and computers in human language or natural language, as it is easier than to communicate using machine language.

Both spoken and written data are dealt with by natural language processing. Though the activities involving data in the form of text i.e., written data is more universal in NLP tasks, raw text data is typically not useable. All such data must be first reshaped into useable format to be fed to the machine or program as an input. The field of NLP encompasses activities like automatic text generation, text analysis, speech recognition, data analysis, smart assistants, and language translation, etc.

The two primary sections of Natural Language Processing are NLU (Natural Language Understanding), often known as Linguistics, and NLG (Natural Language Generation), which come about the work of comprehending and text construction. The scientific study of human language is called linguistic [2][3] or Natural Language Understanding, which is concerned with the several aspects of a language[4] such as sound of speech(phonetics), rules about structure of sentence(syntax), meaning(semantics), structure of words(morphology) and understanding of social context contribution(pragmatics)[5] etc, as shown in figure 2. However, on the other hand, NLG is a software method that creates human language in return of a particular process which may include phrases, sentences, words etc. NLG is a subfield of computational linguistics and AI that deals with developing systems that can construct human language content - such as English, Hindi, French or some other languages - from some fundamental representation of non-linguistic data [6].

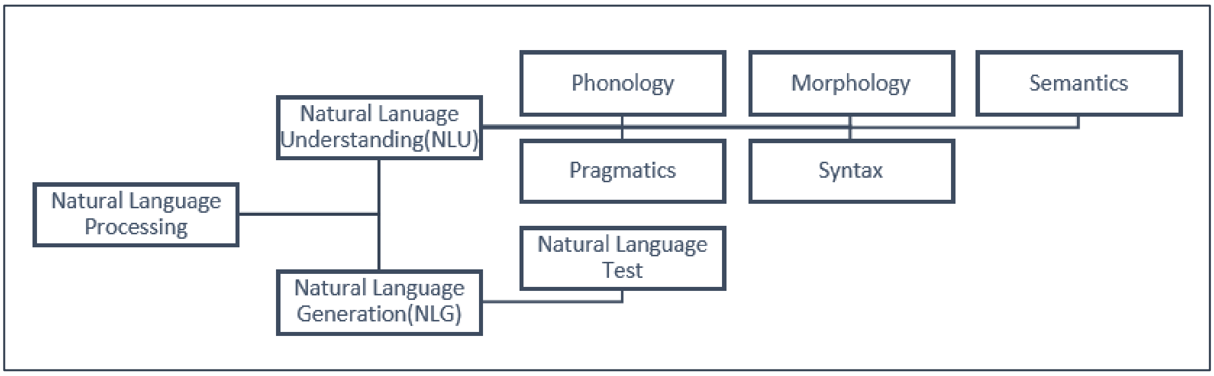

Natural Language Processing
Natural Lanuage Understanding(NLU)
Natural Language Generation(NLG)
Phonology
Morphology
Semantics
Pragmatics
Syntax
Natural Language Test

Fig. 2 NLP Classification

In the subject of natural language processing, numerous ideas and techniques address the challenge of using natural language to communicate with computers. Recognizing the ambiguity of words is a critical task and can be resolved by understanding the entire text and hence the accuracy of the output is certainly based on the algorithm used. Hence, in this following research, a comparative study among various algorithms used in natural language processing (particularly classifiers) is done to know their efficiency on ‘Language Detection’, i.e., an application of NLP which has been mentioned in coming sections.

II. LITERATURE REVIEW

According to Charniak (1995) [7], using straightforward statistical techniques, it is possible to tag a word with its part of speech with 90% accuracy. Jelinek (1999) [8] is a commonly referenced author when statistical methods usage in NLP is concerned, particularly in voice refining.

According to Liddy (2005) [9] and Feldman (1999), understanding natural languages requires being competent in discriminating between the seven interconnected measures employed by people to decipher spoken or written language.

According to Kongthon (2009), an online tax system will be implemented using natural language processing and AI. This implementation was used to demonstrate the viability of using text analysis and NLP in AI to safeguard the future.

In his research, Giuseppe Di Guida (2020) [10] examined the futuristic formal and numerical necessities for textual translation theories, methodologies, and tools to support information modelling and project management processes.

In their research, Diksha Khurana and Aditya Koli (2022) [11], discusses and differentiates the phases of NLP and NLG along with the historical progression of natural language processing, its trends, applications, and challenges.

Cyril Joe Baby and Faizan Ayyub Khan [12] suggests a web application that would allow users to remotely operate fans, lights, and other electrical devices. The first feature of the chatbot algorithm is that it enables customers to enter data so they can govern the action of electronic equipment at home, which is one of the online application's key benefits. Natural language processing methods are used to process the chatbot's messages.

According to Jeonghee (2003) [13], the way a sentiment analyser functions is by capturing sentiments (opinions) on a specific topic. Sentiment extraction, matter-specific feature term extraction, and association via relationship analysis make up opinion mining. The opinion/sentiment lexicon and the related pattern database are a couple of language assets used for analysing the opinions. It scans the documents for both fors and against words, then rates them from -5 to +5.

III. NLP TECHNIQUES

To pre-process the text before any further proceedings of natural language processing, a few techniques and steps are used to extract useful and meaningful data from the larger amount of text. Such techniques have been discussed below:

(1) Tokenization

Tokenization is among the most fundamental and straightforward NLP approaches for processing natural language. It is used to break larger sentences or paragraphs into smaller units and words known as ‘tokens’, such as words, numbers or other characters, that can be easily understood by the machine [14]. The computer can count the frequency of particular words and the places in the text where they often appear thanks to this key step. Example- The stars twinkle at night → ‘The’, ‘stars’, ‘twinkle’, ‘at’, ‘night’.

(2) Segmentation

The process of breaking down written text into understandable units known as ‘segments’, such as words, sentences, or themes, is known as segmentation [15]. It is used by the human mind to process/comprehend the text and also by machines in a similar way. It is easier for the machine to analyse the segments instead of whole text at once. Segmentation is used for various purposes in natural language processing such as emotion analysis, opinion mining, sentiment mining [16][17], etc.

Example- The sky is clear, it will be sunny today → ‘The sky is clear’, ‘it will be sunny today’.

(3) Removing stop words

Conjunctions(but, and, or, because, hence, etc) that are used for connecting sentences or words, Prepositions(in, at, on, under, etc) that are used for expressing relation to other words or helping verbs(is, am, are, was, were, etc) are called stop words and though they may seem like an important part of human speech, do not help a machine much and therefore, can be removed. This helps the machine to focus on only important words and reduces the unnecessary processing [18].

Example- The sun is shining bright → ‘sun’, ‘shining’, ‘bright’.

(4) Stemming and Lemmatization

The method of stemming [19] involves returning derived words to their original form (finding the word stem of a word). Stemming is a crucial method that works by cutting off the end of the words which is one of the reasons that this method may or may not produce a meaningful term in the end.

Example- blinking/blinked/blinks → ‘blink’.

Lemmatization, on the other hand, involves assembling together the derived forms of word and analysing them as a single word (dictionary form) or lemma [20]. This process involves the grouping of words based on the context unlike stemming which is less accurate than lemmatization and operates without any context.

Example- is/am/are → be(lemma)

(5) Part-of-speech Tagging

A grammatically correct sentence consists of part-of-speech such as subject, verb, object, adjectives, preposition, etc and the process of tagging/identifying those parts of a text is known as part-of-speech tagging or POST. POST can be performed using algorithms which can easily discover known as well as unknown/hidden parts of speech based on the context [21].

```
Example- The      sun     is      shining      bright
          ↓        ↓       ↓        ↓             ↓
      (article) (noun) (verb)  (adjective)  (adverb)
```

(6) Named Entity Recognition

The process of tagging/identifying the ‘named entities’ within the text for further analysis is known as named entity recognition (NER) or named entity tagging. It is a classification of textual information into some predetermined categories such as location, organization, person, value, etc.

Example- **Tata Motors Limited**(organization), a **USD 37 billion** (monetary value) organisation, is headquartered in **Mumbai, India**(location).

(7) TF-IDF

TF-IDF is an acronym for Term Frequency-Inverse Document Frequency. TF is the measure of occurrence of a word in a document [22], which shows the relevancy and importance of the word. The words like ‘is’, ‘of’, ‘to’, etc are present very frequently in a text but are of less importance and hence IDF is used to assign lower weight to such words and increase the weight of less frequent words [23].

$$TF(t,d) = count\ of\ 't'\ in\ d/number\ of\ words\ in\ 'd' \quad (1)$$

$$IDF(t) = N/DF(t) \quad (2)$$

$$TF - IDF = Term\ Frequency\ \times Inverse\ Document\ Frequency \quad (3)$$

where DF(t) = occurrences of t in N documents in (2)

Apart from the above-mentioned pre-processing steps or techniques, there are few other techniques such as word embeddings, topic modelling, keyword extraction, text summarization, text classification, sentiment analysis, etc. These techniques are getting better and more accurate every day and they are used in pre-processing the data for numerous applications of natural language processing as shown in figure 3.

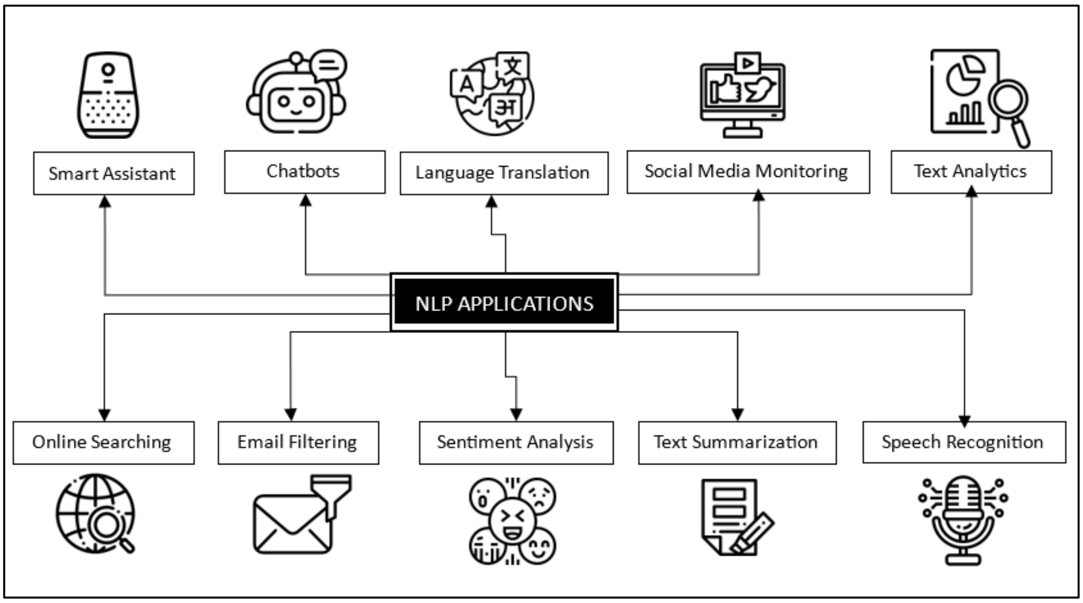


Fig. 3 NLP Applications

IV. OBJECTIVES

- Study and implement natural language processing techniques.
- Study and implement machine learning algorithms and techniques.
- Implement and compare the various machine learning algorithms on an NLP application.
- Study the recent trends and advances in natural language processing.

V. INFORMATION GATHERING

To work on such experiments, a range of wide datasets can be found on the web. The dataset used in this research project is taken from *kaggle* and it contained 10,267 rows of unique data and 2 columns, i.e., text and language. The dataset is small which contains text details for 17 different languages and may be used to build an NLP model for predicting 17 different languages.

VI. PROPOSED SYSTEM

In this research report, the comparative study is done on one of the NLP's applications i.e., 'Language Detection' using machine learning and natural language processing algorithms such as Naïve Bayes Alogrithm, Random Forest Algorithm, Logistic Regression, etc. The training and testing of the data is done using Google Colaboratory. The features from the dataset are visualised using pie-charts, correlation matrix and confusion matrix, etc. This study's primary goal is to check the working of the above mentioned algorithms and their ability to interpret the data in a time bound scenario in the real-world.

VII. METHODOLOGY

The research used python as the coding language for experimental and understanding purposes. This research paper implements various machine leaning and recent NLP algorithms on Google Colaboratory to detect languages. The properties of dataset were understood i.e., counting of unique languages present in the dataset, counting of null-values and duplicate values, removal of redundancy, etc. and we plotted different kinds of charts and plots to visually understand the dataset:

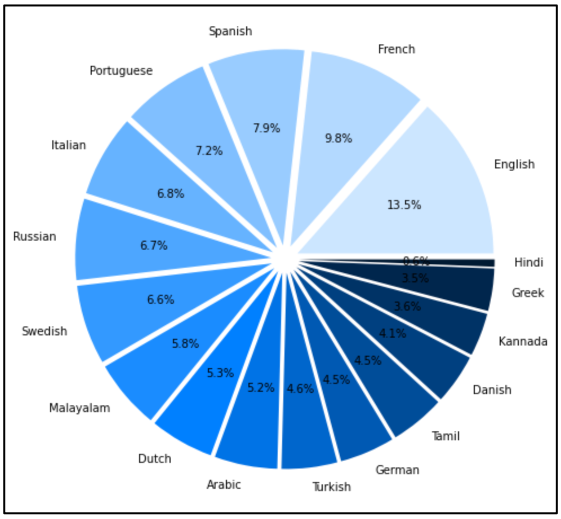


Fig. 4 Language texts pie-chart

The pie-chart shown in figure 4 shows the availability of different language texts in percentage, which makes the comprehension of dataset very easy. The countplot, shown in figure 5, represents the count of texts present in the dataset for the respective languages. It can be seen that the number of English texts is much larger than other languages whereas Hindi texts are the least.

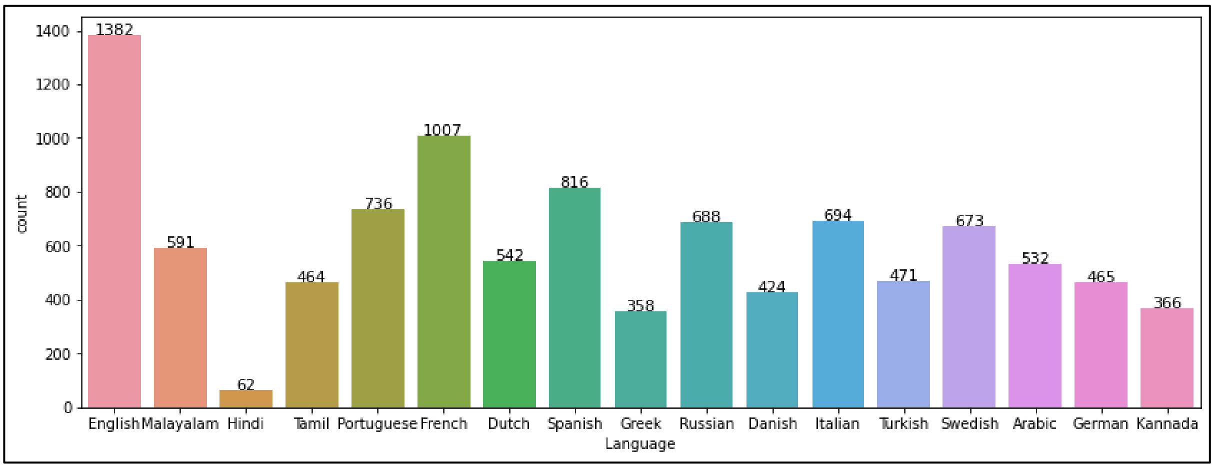


Fig. 5 Count of respective language texts

After understanding the data, pre-processing is performed by cleaning the texts. To clean the text, we eliminate the special characters, numbers, symbols and convert the text to lowercase which is followed by separating the input and output features and encoding the target label class. Feature encoding is done because we can only use numerical values in machine learning models and therefore, it is important to convert the categorical values of the pertinent attributes into numerical. We turn the text into a vector using a count vectorizer depending on the frequency with which each word is used in the text. After formatting and cleaning of dataset, it is broken into training and testing data. The data is then processed using a set of machine learning algorithms. The following algorithms are applied on the dataset:

- Decision Tree
- Multinomial Naïve Bayes
- Logistic Regression
- Random Forest

All above-mentioned algorithms are classifiers and are a part of python's sklearn module. Other python libraries such as Pandas, NumPy, seaborn, matplotlib are also used to create a variety of straightforward and effective tools for machine learning (ML) and data analysis. We have used Google Colaboratory to illustrate the proposed approach and its implementation.

The following explains the above algorithms with their working and output:

(1) Multinomial Naïve Bayes

Among some highly admired classifiers for supervised learning is MNB. It is used for the categorical data exploration. Multinomial naïve bayes is a probabilistic learning algorithm and it is based on Bayes theorem. It is generally used in natural language processing. This method forecasts the tag of a text, such as a newspaper story or email. For a sample of data, it assesses the likelihood of each tag, and then gives out the tag with the highest likelihood in return. The following figure 6 is what we obtain when we plot the confusion matrix between actual and predicted values of MNB classifier on the dataset:

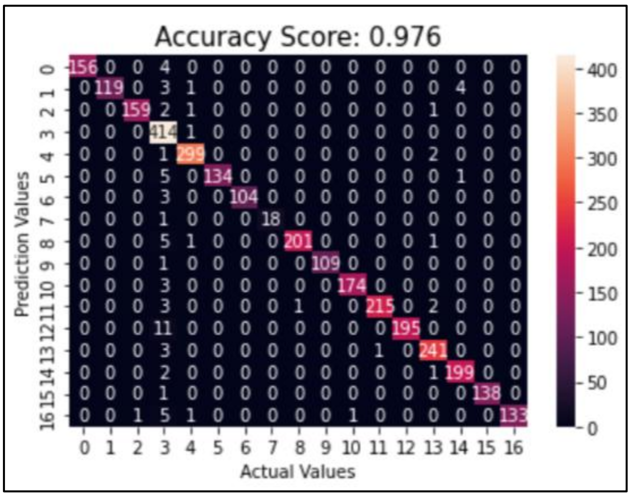
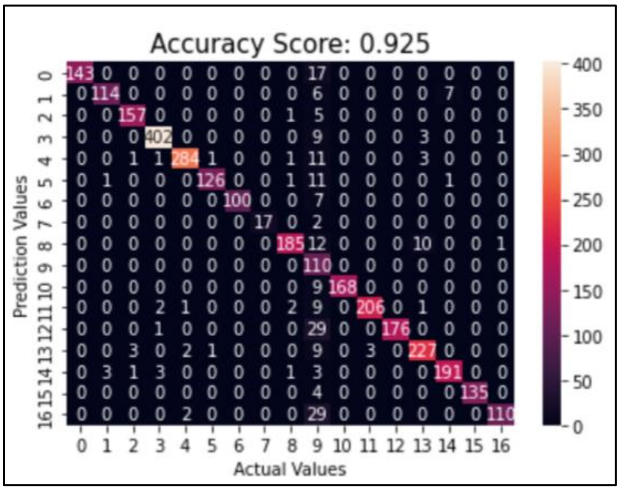


Fig. 6 Confusion matrix of Multinomial Naïve Bayes

(2) Random Forest

This is a famously used classifying algorithm among other supervised machine learning algorithms and can also be used for regression and other decision-making tasks. This classifier works by creating a multilevel tree consisting of several small decision trees using a subset of dataset selected randomly at the time of training. To choose the ultimate prediction, it compiles the votes from various decision trees. The following figure 7 is what we obtain when we plot the confusion matrix between actual and predicted values of random forest classifier on the dataset.

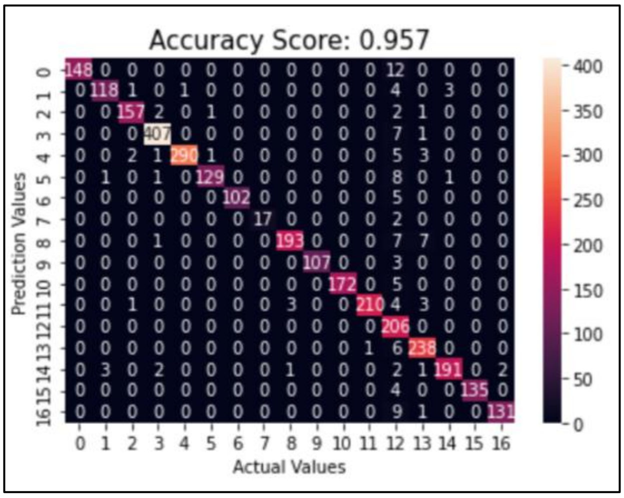


	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	143	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0
1	0	114	0	0	0	0	0	0	0	6	0	0	0	0	7	0	0
2	0	0	157	0	0	0	0	0	1	5	0	0	0	0	0	0	0
3	0	0	0	402	0	0	0	0	0	9	0	0	0	3	0	0	1
4	0	0	1	1	284	1	0	0	1	11	0	0	0	3	0	0	0
5	0	1	0	0	0	126	0	0	1	11	0	0	0	0	1	0	0
6	0	0	0	0	0	0	100	0	0	7	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	17	0	2	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	185	12	0	0	0	10	0	0	1
9	0	0	0	0	0	0	0	0	0	110	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	9	168	0	0	0	0	0	0
11	0	0	0	2	1	0	0	0	2	9	0	206	0	1	0	0	0
12	0	0	0	1	0	0	0	0	0	29	0	0	176	0	0	0	0
13	0	0	3	0	2	1	0	0	0	9	0	3	0	227	0	0	0
14	0	3	1	3	0	0	0	0	1	3	0	0	0	0	191	0	0
15	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	135	0
16	0	0	0	0	2	0	0	0	0	29	0	0	0	0	0	0	110

Fig. 7 Confusion matrix of Random Forest

(3) Logistic Regression

This is also a supervised machine learning algorithm used for the purpose classification. It is based on probabilistic learning method used to predict the probability of target variable. It classifies the variable into two 0(failure/no) or 1(success/yes). As one of the fundamental machine learning algorithms, it may be used to classify a variety of problems, such as cancer diagnosis, diabetes prediction and spam detection, etc. The following figure 8 is what we obtain when we plot the confusion matrix between actual and predicted values of logistic regression classifier on the dataset.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	148	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0
1	0	118	1	0	1	0	0	0	0	0	0	0	4	0	3	0	0
2	0	0	157	2	0	1	0	0	0	0	0	0	2	1	0	0	0
3	0	0	0	407	0	0	0	0	0	0	0	0	7	1	0	0	0
4	0	0	2	1	290	1	0	0	0	0	0	0	5	3	0	0	0
5	0	1	0	1	0	129	0	0	0	0	0	0	8	0	1	0	0
6	0	0	0	0	0	0	102	0	0	0	0	0	5	0	0	0	0
7	0	0	0	0	0	0	0	17	0	0	0	0	2	0	0	0	0
8	0	0	0	1	0	0	0	0	193	0	0	0	7	7	0	0	0
9	0	0	0	0	0	0	0	0	0	107	0	0	3	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	172	0	5	0	0	0	0
11	0	0	1	0	0	0	0	0	3	0	0	210	4	3	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	206	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	1	6	238	0	0	0
14	0	3	0	2	0	0	0	0	1	0	0	0	2	1	191	0	2
15	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	135	0
16	0	0	0	0	0	0	0	0	0	0	0	0	9	1	0	0	131

Fig. 8 Confusion matrix of Logistic Regression

(4) Decision Tree

Decision tree is a popular model used for classification and prediction. It is a tree-like structure having leaf nodes and internal nodes. The leaf-nodes are the end-nodes that specifies the class whereas, the internal nodes are the test nodes based on which the splitting of the node happens. The branches of the tree represent the decision. The following figure 9 is what we obtain when we plot the confusion matrix between actual and predicted values of logistic regression classifier on the dataset.

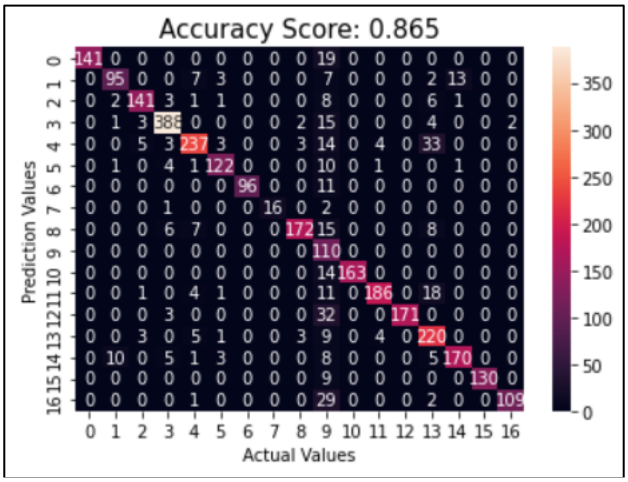


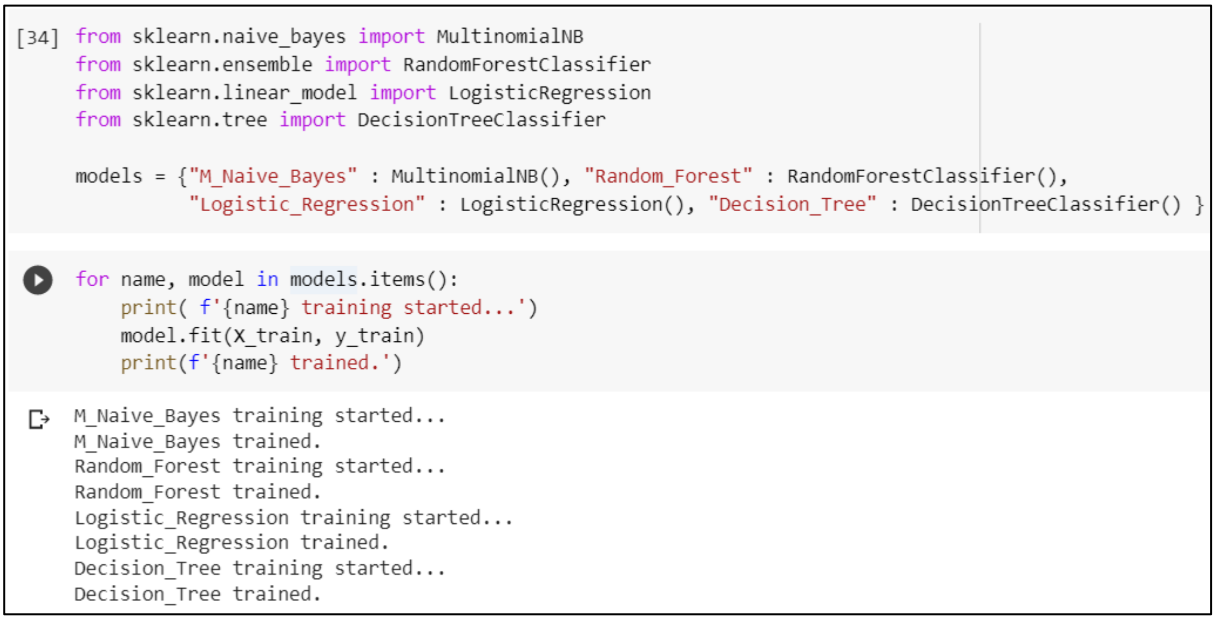
Fig. 9 Confusion matrix of Decision Tree

For comparison and testing purposes, we will compare the results of the algorithms in the next section of this research to find the best classifying algorithm among the mentioned algorithms.

Apart from this, we have performed the analysis of the literature which is already in existence to gather a detailed knowledge to understand NLP and NLG. The literature that was already in existence has been extracted with the aid of keywords after taking into consideration various research papers and earlier analyses. Information from those literatures was extracted, and when it was taken out, it was verified for accuracy and dependability using the right sources. Details were included in the paper and irrelevant information was eliminated after the accuracy test. Following the incorporation of the pertinent data into the research, a second examination of this issue was completed.

VIII. IMPLEMENTATION

This research proposes the idea to implement ML and advance NLP algorithms for language detection. Therefore, we have implemented the algorithms such as random forest, naïve bayes, logistic regression and decision tree to make a comparison between their efficiency in detecting the languages. The pseudocode for the above classifiers is as follows, as shown in figure 10(a) and 10(b) :

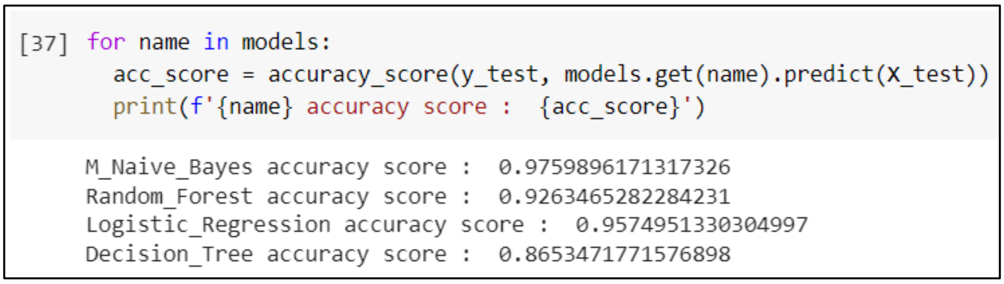


```
[34] from sklearn.naive_bayes import MultinomialNB
     from sklearn.ensemble import RandomForestClassifier
     from sklearn.linear_model import LogisticRegression
     from sklearn.tree import DecisionTreeClassifier

     models = {"M_Naive_Bayes" : MultinomialNB(), "Random_Forest" : RandomForestClassifier(),
               "Logistic_Regression" : LogisticRegression(), "Decision_Tree" : DecisionTreeClassifier() }
```

```
for name, model in models.items():
    print( f'{name} training started...')
    model.fit(X_train, y_train)
    print(f'{name} trained.')
```

```
M_Naive_Bayes training started...
M_Naive_Bayes trained.
Random_Forest training started...
Random_Forest trained.
Logistic_Regression training started...
Logistic_Regression trained.
Decision_Tree training started...
Decision_Tree trained.
```

(a)



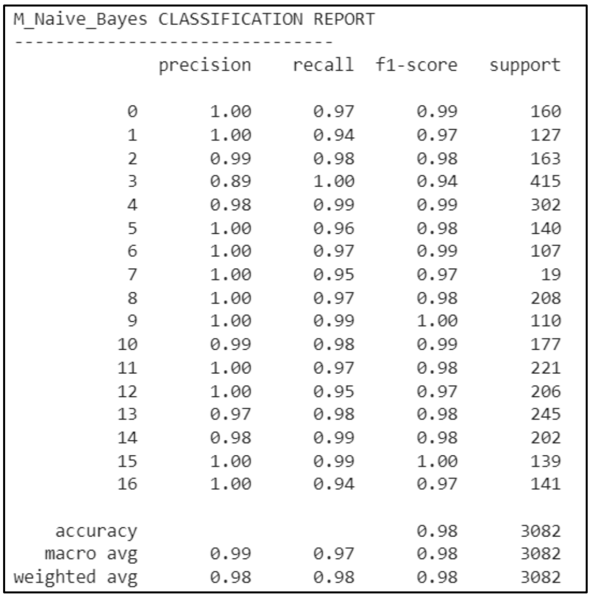
```
[37] for name in models:
       acc_score = accuracy_score(y_test, models.get(name).predict(X_test))
       print(f'{name} accuracy score :  {acc_score}')

     M_Naive_Bayes accuracy score :  0.9759896171317326
     Random_Forest accuracy score :  0.9263465282284231
     Logistic_Regression accuracy score :  0.9574951330304997
     Decision_Tree accuracy score :  0.8653471771576898
```

(b)

Fig. 10 Various (a) Model training and (b) Accuracy scores

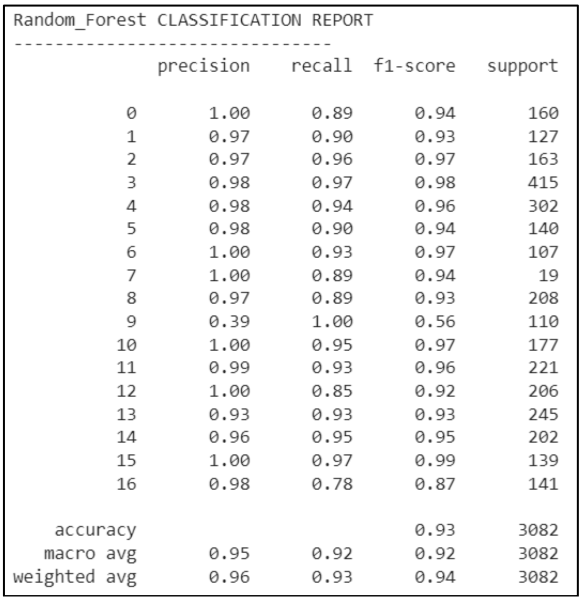
IX. RESULTS

The algorithms are compared based on accuracy score, confusion matrices, precision, recall, etc. These outcomes are provided in the output along with the classification report for the respective algorithms as follows, where the different numbers represent the different languages present in the dataset.



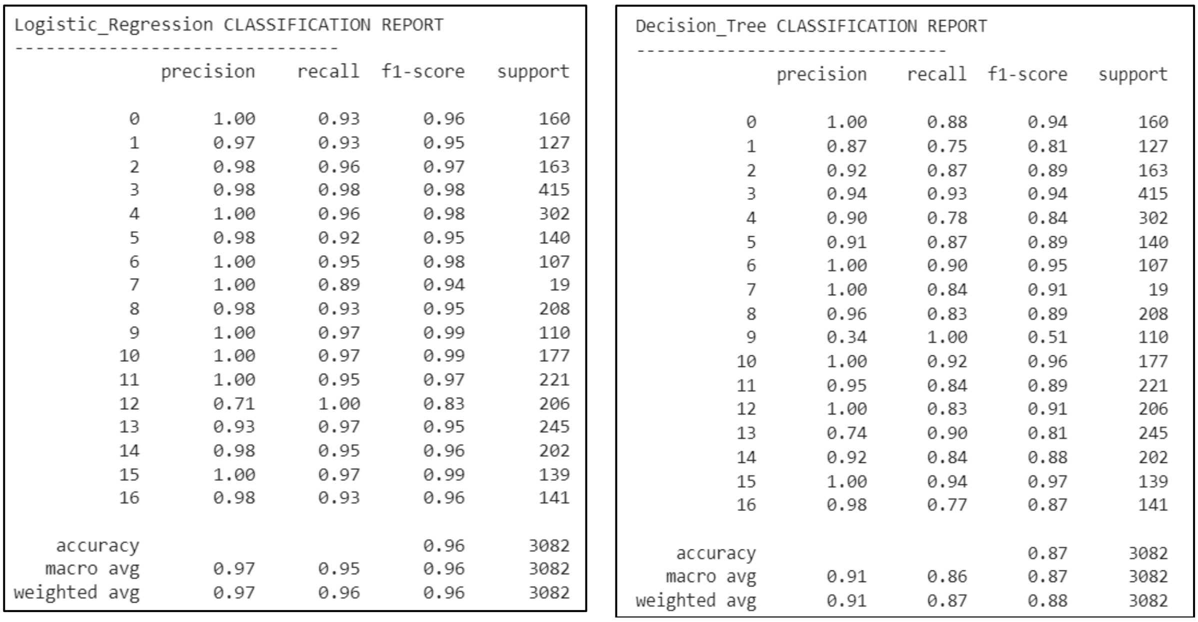
```
M_Naive_Bayes CLASSIFICATION REPORT
------------------------------
              precision    recall  f1-score   support

           0       1.00      0.97      0.99       160
           1       1.00      0.94      0.97       127
           2       0.99      0.98      0.98       163
           3       0.89      1.00      0.94       415
           4       0.98      0.99      0.99       302
           5       1.00      0.96      0.98       140
           6       1.00      0.97      0.99       107
           7       1.00      0.95      0.97        19
           8       1.00      0.97      0.98       208
           9       1.00      0.99      1.00       110
          10       0.99      0.98      0.99       177
          11       1.00      0.97      0.98       221
          12       1.00      0.95      0.97       206
          13       0.97      0.98      0.98       245
          14       0.98      0.99      0.98       202
          15       1.00      0.99      1.00       139
          16       1.00      0.94      0.97       141

    accuracy                           0.98      3082
   macro avg       0.99      0.97      0.98      3082
weighted avg       0.98      0.98      0.98      3082
```

(a)



```
Random_Forest CLASSIFICATION REPORT
------------------------------
              precision    recall  f1-score   support

           0       1.00      0.89      0.94       160
           1       0.97      0.90      0.93       127
           2       0.97      0.96      0.97       163
           3       0.98      0.97      0.98       415
           4       0.98      0.94      0.96       302
           5       0.98      0.90      0.94       140
           6       1.00      0.93      0.97       107
           7       1.00      0.89      0.94        19
           8       0.97      0.89      0.93       208
           9       0.39      1.00      0.56       110
          10       1.00      0.95      0.97       177
          11       0.99      0.93      0.96       221
          12       1.00      0.85      0.92       206
          13       0.93      0.93      0.93       245
          14       0.96      0.95      0.95       202
          15       1.00      0.97      0.99       139
          16       0.98      0.78      0.87       141

    accuracy                           0.93      3082
   macro avg       0.95      0.92      0.92      3082
weighted avg       0.96      0.93      0.94      3082
```

(b)



```
Logistic_Regression CLASSIFICATION REPORT
-------------------------------
              precision    recall  f1-score   support

           0       1.00      0.93      0.96       160
           1       0.97      0.93      0.95       127
           2       0.98      0.96      0.97       163
           3       0.98      0.98      0.98       415
           4       1.00      0.96      0.98       302
           5       0.98      0.92      0.95       140
           6       1.00      0.95      0.98       107
           7       1.00      0.89      0.94        19
           8       0.98      0.93      0.95       208
           9       1.00      0.97      0.99       110
          10       1.00      0.97      0.99       177
          11       1.00      0.95      0.97       221
          12       0.71      1.00      0.83       206
          13       0.93      0.97      0.95       245
          14       0.98      0.95      0.96       202
          15       1.00      0.97      0.99       139
          16       0.98      0.93      0.96       141

    accuracy                           0.96      3082
   macro avg       0.97      0.95      0.96      3082
weighted avg       0.97      0.96      0.96      3082
```

(c)

```
Decision_Tree CLASSIFICATION REPORT
------------------------------
              precision    recall  f1-score   support

           0       1.00      0.88      0.94       160
           1       0.87      0.75      0.81       127
           2       0.92      0.87      0.89       163
           3       0.94      0.93      0.94       415
           4       0.90      0.78      0.84       302
           5       0.91      0.87      0.89       140
           6       1.00      0.90      0.95       107
           7       1.00      0.84      0.91        19
           8       0.96      0.83      0.89       208
           9       0.34      1.00      0.51       110
          10       1.00      0.92      0.96       177
          11       0.95      0.84      0.89       221
          12       1.00      0.83      0.91       206
          13       0.74      0.90      0.81       245
          14       0.92      0.84      0.88       202
          15       1.00      0.94      0.97       139
          16       0.98      0.77      0.87       141

    accuracy                           0.87      3082
   macro avg       0.91      0.86      0.87      3082
weighted avg       0.91      0.87      0.88      3082
```

(d)

Fig. 11 Classification report of (a) Multinomial Naïve Bayes (b) Random Forest (c) Logistic Regression and (d) Decision Tree on Language Detection

It can be seen in the above classification reports that precision of MNB classifier(figure 11a) is the highest among the other classifiers whereas Decision tree(figure 11d) shows only 91% precision i.e., lowest than others. Logistic regression and random forest also provided us a good precision score(figure 11b, 11c). Except from these well-known algorithms used in natural language processing, the world has made a highly noticeable advance in the field by introducing several other major algorithms which are discussed in the coming section along with new trends that has made natural language processing easier and more accurate.

X. RECENT TRENDS AND ADVANCES IN NLP

(1) Trends in Natural Language Processing

- **Transfer Learning**

As the name implies, the transfer of the knowledge and learnings obtained from a problem to another problem that is unrelated yet similar is known as transfer learning [24]. Transfer learning is a research problem of machine learning (ML). For example, the knowledge obtained from identifying a bicycle can be applied to identify a motorcycle. In other words, you will just need to make modest tweaks to a pre-trained model rather than spending a lot of money, time, and effort constructing and training a model from start. Some of TL applications are digital recognition, game playing, text classification and medical imaging.

- **Low code tools**

NLP requires a good set of skills such as coding to use libraries, background knowledge of the concerned field and machine learning algorithms, etc. to develop a model but it has been made easier by low-code or no code tools provided by some organizations that enables the non-technical individuals to do out tasks requiring NLP, which were only could be done by the professionals earlier.

- **Language Transformers**

Language transformers are deep learning models that employ the self-attention technique and assign varying weights to the significance of each input parameter. Language transformers use the data in a sequential form and with the help of attention mechanism, it provides the context of input data at any position. Unlike other models, language transformer takes the entire data as the input at once. Transformers debuted in 2017 by Google Brain and has been evolving ever since. The NLP is currently focused on transformers such as GPT (Generative Pre-trained Transformer) [25] and BERT (Bidirectional Encoder Representations from Transformers).

- **Multilingual NLP**

The world has over 7000 languages being spoken but most natural language processors rely on English and hence do not perform efficiently. The construction of NLP models is made possible by the availability of large training datasets in a variety of languages and hence companies such as Google and Facebook are bringing out the multilingual pre-trained models such as XLM-R and M2M-100, that work as efficiently as they work on monolingual datasets.

- **Reinforcement Learning**

The process of learning through exploring the environment and getting rewards and penalties as feedback to an action, is reinforcement learning. It allows NLP models to learn actions that increase the likelihood of a successful outcome and improve its performance with the help of reward-based training iteration series. RL can be used to accelerate processes like translation, summarization, etc. in NLP.

(2) Advances In Natural Language Processing

- **BERT**

The acronym stands for Bidirectional Encoder Representations from Transformers. This model was developed by Google AI in 2018 and is a well-known language representation model. BERT is based on transformer architecture. A transformer works on a self-attention mechanism [26]. This attention mechanism enables us to make analogous connections within a single sentence and hence, the transformer is a unique approach to sequence-to-sequence difficulties in NLP that deftly handles long-range dependencies. It completely counts on self-attention and does not calculate its input and output representations using sequence aligned RNNs or convolution. The model is intended for pre-training the deep bidirectional representations. This model is cutting-edge for 11 NLP tasks and hence it has been researched by organization such as Facebook and Academia.

- **Google's TransformerXL**

Developed by Google AI in 2019, this model is a novel neural architecture which even defeated the famous BERT in language modelling. This model permits the learning dependency with more flexibility without interfering with temporal coherence. The original transformers had context fragmentation problem which was fixed by this model.

- **XLNet**

The model was proposed by Google AI Brain team and Carnegie Mellon University's research team in June 2019. It resolves the issues related to BERT using its autoregressive formulation and amalgamates the TransformerXL ideas, combining the top qualities of both. In XLNet, the language modelling is done using the auto-regressive methods unlike BERT which uses auto-encoding methods.

- **RoBERTa**

The acronym stands for Robustly Optimized BERT Approach. This model was developed by Facebook AI researchers as the name implies, RoBERTa is a variation of BERT. RoBERTa and BERT differ significantly from each other in that RoBERTa was learned using a larger dataset (160GB of text) and a more efficient training method. Like BERT, it is also a transformer model and uses the self-attention mechanism.

Apart from the above mentioned, there has been many other models that helped in the advancement of NLP. For example, ULMFiT (Universal Language Model Fine- Tuning), OpenAI's GPT-2, PyTorch-Transformers, Baidu's ERNIE (Enhanced Representation through kNowledge IntEgration), etc.

XI. CONCLUSION

This paper has implemented and listed a few machine learning and natural language processing algorithms that are used for language detection and for other text classification purposes. The research also mentioned the steps and techniques used for pre-processing in natural language processing along with their examples. This research presented a comparative study of the algorithms and how they can be used to get better output, along with their implementation and outcomes. Famous classifiers were used for the experimental purposes such as random forest, naïve bayes, logistic regression and decision tree. While we got the average accuracy of 95.7% with logistic regression, 92.5% with random forest, decision tree only gave us an accuracy of 86.5%. The highest accuracy was given by naïve bayes, i.e., 97.5%. This accuracy can be certainly increased if more and efficient data is fed to the algorithms. Finally, the recent trends and advances in the field of natural language processing are discussed in a detailed manner.

XII. FUTURE SCOPE

From this research we can see that, although we may not have reached a higher accuracy than 98% approximately but it is much better than other systems and can perform more efficiently with more efficient data in given time. There is always a room for betterment in a system. The amount and quality of dataset decides the precision and accuracy of the results. Therefore, more data will provide better output. Besides this, more algorithms can be used together to build a better model. Today, transformers are trending, and they are being used by biggest organizations in their search engines, chatbot systems, text classification, etc. The results are also pretty accurate and close due to their self-attention mechanism. Big and renowned organizations are working on transformers since 2018 and gaining so much success. It has changed the language processing by giving better results than RNN or convolution network algorithms and continues to do so. Enhanced algorithms such as the BERT, XLNet, RoBERTa, etc are only the steppingstones to a far1 better future.

XIII. REFERENCES

[1] N. Chomsky, *Aspects of the Theory of Syntax*. The MIT Press, 1965. doi: 10.1604/9780262530071.

[2] M. A. K. Halliday and J. Webster, *On Language and Linguistics: Volume 3*. Continuum, 2006. doi: 10.1604/9780826488244.

[3] "What is Linguistics? | Linguistic Society of America," *What is Linguistics? | Linguistic Society of America*. https://www.linguisticsociety.org/what-linguistics

[4] D. Crystal, *Clinical Linguistics*, 1st ed. Springer Vienna, 2013. doi: 10.1007/978-3-7091-4001-7.

[5] A. Akmajian, R. A. Demers, A. K. Farmer, and R. M. Harnish, *Linguistics: An Introduction to Language and Communication*. 6th ed. The MIT Press, 2010.

[6] E. Reiter and R. Dale, "Building applied natural language generation systems," *Natural Language Engineering*, vol. 3, no. 1, pp. 57–87, Mar. 1997, doi: 10.1017/s1351324997001502.

[7] E. Charniak, "Natural language learning," *ACM Computing Surveys*, vol. 27, no. 3, pp. 317–319, Sep. 1995, doi: 10.1145/212094.212108.

[8] F. Jelinek and D. X. Sun, "Statistical Methods for Speech Recognition," *Journal of the American Statistical Association*, vol. 94, no. 446, p. 650, Jun. 1999, doi: 10.2307/2670189.

[9] E. D. Liddy, "Enhanced Text Retrieval Using Natural Language Processing," *Bulletin of the American Society for Information Science and Technology*, vol. 24, no. 4, pp. 14–16, Jan. 2005, doi: 10.1002/bult.91.

[10] G. Giuda, M. Locatelli, M. Schievano and E. Seghezzi, "Natural Language Processing for Information and Project Management," *Digital Transformation of the Design, Construction and Management Processes of the Built Environment*, Springer, Cham, 2020. doi: 10.1007/978-3-030-33570-0_9.

[11] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimedia Tools and Applications*, vol. 82, no. 3, pp. 3713–3744, Jul. 2022, doi: 10.1007/s11042-022-13428-4.

[12] C. J. Baby, F. A. Khan, and J. N. Swathi, "Home automation using IoT and a chatbot using natural language processing," *2017 Innovations in Power and Advanced Computing Technologies(i-PACT)*, Vellore, India, 2017, pp. 1-6, doi: 10.1109/IPACT.2017.8245185..

[13] J. Yi, T. Nasukawa, R. Bunescu and W. Niblack, "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques," *2003 IEEE International Conference on Data Mining(ICDM)*, 2003, pp. 427-434, doi: 10.1109/ICDM.2003.1250949.

[14] J. Webster and C. Kit, "Tokenization as the initial phase in NLP," 1992 *14th Conference on Computational Linguistics*, Aug. 1992, pp. 1106- 1110, doi: 10.3115/992424.992434

[15] J. B. Lovins, "Development of a Stemming Algorithm," *Mechanical Translation and Computational Linguistics*, Vol. 11, No. 1-2, 1968, pp. 22-31.

[16] A. A. Hakim, A. Erwin, K. I. Eng, M. Galinium and W. Muliady, "Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach," *2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE)*, Yogyakarta, Indonesia, 2014, pp. 1-4, doi: 10.1109/ICITEED.2014.7007894.

[17] J. West, D. Ventura and S. Warnick, “Spring Research Presentation: A Theoretical Foundation for Inductive Transfer,” Brigham Young University, College of Physical and Mathematical Sciences, 2007

[18] S. Sarica and J. Luo, “Stopwords in technical language processing,” *PLOS ONE*, vol. 16, no. 8, p. e0254937, Aug. 2021, doi: 10.1371/journal.pone.0254937.

[19] Deepak, K. Visweswariah, N. Wiratunga and S. Sani, “Two-part segmentation of text documents,” *2012 21st ACM International Conference on Information Knowledge Management*, 2012, pp. 793, doi: 10.1145/2396761.2396862

[20] Y. Gao, L. Zhou, Y. Zhang, C. Xing, Y. Sun and X. Zhu, "Sentiment classification for stock news," *2010 5th International Conference on Pervasive Computing and Applications*, Maribor, Slovenia, 2010, pp. 99-104, doi: 10.1109/ICPCA.2010.5704082.

[21] H. Xia, M. Tao and Y. Wang, "Sentiment text classification of customers reviews on the Web based on SVM," *2010 Sixth International Conference on Natural Computation*, Yantai, China, 2010, pp. 3633-3637, doi: 10.1109/ICNC.2010.5584077.

[22] T. Müller, R. Cotterell, A. Fraser and H. Schütze, “Joint Lemmatization and Morphological Tagging with LEMMING,” *2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2268–2274, doi:10.18653/v1/D15-1272

[23] A. Søgaard, “Simple semi-supervised training of part-of-speech taggers,” *ACL 2010 Conference Short Papers*, 2010, pp. 205–208

[24] A. Aizawa, “An information-theoretic perspective of tf–idf measures,” *Information Processing & Management*, vol. 39, no. 1, pp. 45–65, Jan. 2003, doi: 10.1016/s0306-4573(02)00021-3.

[25] "Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing," *Google AI Blog*, Retrieved 2023-03-25

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser and I. Polosukhin, “Attention is All you Need,” *NIPS*, Jun. 2017